Success Story

Machine Learning and Feature Selection for Population Health

HEALTHCARE ORGANIZATION

Integrated Delivery System

PRODUCTS

- Health Catalyst® Data Operating System (DOS™) Platform

SERVICES

- Professional Services

EXECUTIVE SUMMARY

To improve population health, Christiana Care Health System (CCHS) had used a machine learning model to inform population segmentation. The initial model used “black box” algorithms to predict risk that care managers didn’t have input on or understand. As a result, the care managers didn’t trust or use the model, a common problem with machine learning algorithms.

CCHS leaders and experts wanted a model they could use to predict 90-day inpatient admission because they felt that it was a meaningful outcome for care management intervention. CCHS built a new machine learning model using an analytics platform that had a feature selection process to build the simplest model possible—and AI insight tools for selecting the best model, setting triggers for action, and explaining how the model worked. Results include:

- Feature selection reduced the model complexity from 236 data features to just 16 data features (7 percent of the original data set).
- Both models, the one with 236 data features and the one with 16 data features, had an AUROC of 0.78 and an AUPR of 0.15, suggesting no degradation of predictive performance due to the lower number of features selected.
- Clinical experts feel like the selected features agreed with their experience and expectations. As a result, CCHS care managers have confidence in the predictive model, and they are successfully using the output of the machine learning tool to engage with an average of 857 distinct members each week, completing more than 2,520 tasks for those members.
- Feature selection and the other analytics platform tools are broadly applicable to any dataset, which will enable CCHS to use them to develop more models in the future.

> When healthcare first started talking about population health, the focus was on care management. Now, the focus in population health is on data and analytics, these are core to successfully caring for populations.
>
> Terri Steinberg, MD, MBA
> Chief Health Information Officer and Vice President for Population Health Informatics
> Christiana Care Health System

MANAGING POPULATIONS WITH CHRONIC CONDITIONS

Ninety percent of the $3.3 trillion spent in the United States annually for healthcare is for people with chronic and mental health conditions.[1] This makes patient populations with chronic conditions the most frequent and expensive users of care.

Addressing these populations, value-based care models have expanded with the intention of lowering healthcare costs by encouraging better management of patients with existing conditions to prevent hospitalizations, avoid complications, and improve outcomes.

Delaware-based CCHS's mission is to serve its neighbors as expert, caring partners in their health. The health system consistently creates innovative, effective, affordable systems of care. It sought to improve population health in its community and lower healthcare costs by incorporating value-based care models.

UNTRUSTED ALGORITHMS IMPACT USAGE

CCHS developed a patient-centered and clinician-led service for enhanced care coordination to support providers and care managers in helping patients achieve better clinical outcomes at a lower cost. The healthcare organization had used a machine learning model for population segmentation and care management, incorporating results and admission-discharge-transfer data to generate a numeric risk score that was rolled into risk bands for presenting to care managers in their native workflow.

The model used "black box" algorithms that the care managers didn't have input on and didn't understand. Hence, they were not trusted or used by many individuals, which is a common problem with machine learning algorithms.

As CCHS's processes matured, the organization sought to improve its risk prioritization, enabling it to better identify patients who are likely to have a hospital admission within the next 90-days. This helps care managers engage with patients to ensure they receive the right care, at the right time, in the right place.

> Successful population health management requires the segmentation of populations and being able to predict individuals with the highest risk and those with rising risk. Machine learning models are needed to accurately make these predictions. However, it's not enough to build machine learning models. AI insights are needed to build the simplest machine learning models possible, and the models need to be understood and trusted by clinicians if we want them to use the output from the models.
>
> Terri Steinberg, MD, MBA
> Chief Health Information Officer and Vice President for Population Health Informatics
> Christiana Care Health System

Accessing clinical data sets is often difficult and expensive, particularly when those sets are derived from different health systems. This challenge increased when CCHS expanded beyond its care management services outside of the state of Delaware, where it had a custom integration with the state's health information exchange (HIE). CCHS wanted certainty it was using the best data to improve population health as its care teams were already inundated with information, and care managers were skeptical of the existing machine learning model.

CCHS leaders sought to answer some key questions:

- Would more data help or be overwhelming?
- What kind of data are needed to improve the risk score methodology for population segmentation?
- When data are added, do risk and outcomes predictions become more accurate?

Before undergoing the challenging process of incorporating additional, expensive data sets to its risk prediction model, CCHS needed to better understand the benefit of adding clinical data to its claims information.

DATA-DRIVEN APPROACH TO IMPROVE POPULATION HEALTH

To answer these key questions surrounding additional data, CCHS partnered with Health Catalyst to leverage the Health Catalyst® Data Operating System (DOS™) and its data science team to develop a new machine learning model to predict hospital admission within the next 90-days. CCHS saw this as a way to use machine learning in a pragmatic way.

Creating a machine learning model to predict hospital admissions

In developing the machine learning model, CCHS analyzed data from more than 100,000 individuals, including those enrolled in a Medicare Accountable Care Organization (ACO), CCHS patients, members of direct-to-employer contracts, and payer/care management partnerships.

The previous approach to segmenting populations using risk scores would have required analysis of up to 236 columns of data “features.” This complex approach is difficult to develop and maintain, prompting CCHS to move away from clinical data, instead utilizing claims and enrollment data for initial models.

The data cohort included various payer contracts and other populations. Each group, within the 100,000-member unit, had vastly different data characteristics, with disparities in healthcare utilization, member age, demographics, diseases, and socioeconomic status. During data analysis, the team worked to understand why some data were missing, as these reasons dictate how “missingness” is handled when building and selecting the features in the model.

Data acquisition from HIEs was also a challenge. During the lookback period, which is 30-days for labs and six-months for other data, there can be many clinical events per member. This translates into millions of rows for hundreds of columns of data that need to be loaded and aggregated initially, and then appended to, updated, and aggregated going forward. Further, clinical events in HIEs are often not mapped to common nomenclature/terminology (e.g., lab results mapped to LOINC), making it difficult to supplement missing data or combine data that comes from different sources.

Establishing a model that enables ease of future use

To simplify future data collection, CCHS identified those elements that were the greatest contributors to risk stratification predictive performance. To identify the incremental benefit of adding clinical data, the team used machine learning in a process known as “feature selection” to test the incremental predictive contribution of each of the 236 data features by iteratively evaluating whether each feature, alone or in combination, had a material impact on the overall predictive performance of the model.

A feature was then included in the model if its inclusion resulted in a 0.005 incremental increase in the AUROC curve or the AUPR curve, the established performance measures used to evaluate and compare the accuracy, precision, and sensitivity of analytical models. To ensure that no degradation of predictive performance resulted from feature selection, model performance after feature selection was compared to model performance before feature selection.

Once the 90-day-admission predictive model was created, CCHS care managers were trained on the model, helping them understand and trust the risk score provided and apply it in caring for their patients.

RESULTS

By understanding the implication of all population health data feeding its machine learning model, CCHS was able to establish trust among care managers and deliver results, including:

- Feature selection reduced the model complexity from 236 data features to just 16 data features (7 percent of the original data set).
- Both models, the one with 236 data features and the one with 16 data features, had an AUROC of 0.78 and an AUPR of 0.15, suggesting no degradation of predictive performance due to the lower number of features selected.
- Feature selection and the other analytics platform tools are broadly applicable to any dataset, which will enable CCHS to use them to develop more models in the future.

The model runs daily, and updates predictions for about 158,000 active patients as new information about them becomes available. The score is then presented within CCHS's care management software application in a simplified form: the highest-risk patients are provided as a list that can be filtered by the top 5 percent, 10 percent, or 15 percent of members to engage.

Clinical experts feel like the selected features agreed with their experience and expectations. As a result, CCHS care managers have confidence in the predictive model, and they are successfully using the output of the machine learning tool to engage with an average of 857 distinct members each week, completing more than 2,520 tasks for those members.

WHAT'S NEXT

CCHS will continue to pursue opportunities to leverage analytics and data science for improving population health outcomes. The health system will also continue to ensure that every person gets the right care, at the right time, in the right place.

REFERENCES

1. Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion. (2019). *Health and economic costs of chronic diseases.* Retrieved from https://www.cdc.gov/chronicdisease/about/costs/index.htm

ABOUT HEALTH CATALYST

Health Catalyst is a leading provider of data and analytics technology and services to healthcare organizations, committed to being the catalyst for massive, measurable, data-informed healthcare improvement. Our customers leverage our cloud-based data platform—powered by data from more than 100 million patient records, and encompassing trillions of facts—as well as our analytics software and professional services expertise to make data-informed decisions and realize measurable clinical, financial, and operational improvements. We envision a future in which all healthcare decisions are data informed. Learn more at www.healthcatalyst.com.

Visit www.healthcatalyst.com, and follow us on Twitter, LinkedIn, and Facebook.